Relative and objective, on balance

Bialek, Max

A

SHANNON'S BSA

In this appendix, I provide a reconstruction of a classic theorem from Shannon (1948), but newly motivated to be about induction friendliness and the BSA. The theorem and its setup is worthy of being included here in full.[1]

> We have represented a discrete information source as a Markoff process. Can we define a quantity which will measure, in some sense, how much information is "produced" by such a process, or better, at what rate information is produced?
>
> Suppose we have a set of possible events whose probabilities of occurrence are $p_1, p_2, ..., p_n$. These probabilities are known but that is all we know concerning which event will occur. Can we find a measure of how much "choice" is involved in the selection of the event or of how uncertain we are of the outcome?
>
> If there is such a measure, say $H(p_1, p_2, ..., p_n)$, it is reasonable to require of it the following properties:
>
> 1. H should be continuous in the p_i.
> 2. If all the p_i are equal, $p_i = 1/n$, then H should be a monotonic increasing function of n. With equally likely events there is more choice, or uncertainty, when there are more possible events.
> 3. If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H. The meaning of this is illustrated in Fig. [2].
>
>   At the left we have three possibilities $p_1 = \frac{1}{2}$, $p_2 = \frac{1}{3}$, $p_3 = \frac{1}{6}$. On the right we first choose between two possibilities each with probability $\frac{1}{2}$, and if the second occurs make another choice with probabilities $\frac{2}{3}$, $\frac{1}{3}$. The

1 The proof, however, I leave for the most committed of readers to find in the second appendix to Shannon (1948).

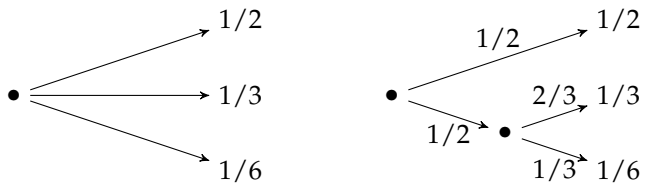


Figure 2: Decomposition of a choice from three possibilities.

final results have the same probabilities as before. We require, in this special case, that

$$H(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}) = H(\frac{1}{2}, \frac{1}{2}) + \frac{1}{2}H(\frac{2}{3}, \frac{1}{3})$$

The coefficient $\frac{1}{2}$ is the weighting factor introduced because this second choice only occurs half the time.

In Appendix 2, the following result is established:

Theorem 2: The only H satisfying the three above assumptions is of the form:

$$H = -K \sum_{i=1}^{n} p_i \log p_i$$

where K is a positive constant.

(Shannon 1948, p. 10)

This is what gives rise to the "uncertainty" interpretation of entropy. Shannon himself notes that this theorem is unnecessary to the minimum average code length interpretation of entropy that is the primary concern of information theory (at least as far as Shannon was concerned in that paper). But it is illuminating.

And it is that illuminating quality that prompts the remaining work of this appendix. I reconstruct this theorem in the context of the BSA, leading to an interpretation of entropy as, roughly, "induction unfriendliness when the present state of the world is known". More importantly, from there we can derive (approximately) MI as a measure of induction friendliness.

I begin by introducing a toy model of the BSA (Section A.1) that parallels Shannon's toy model of a communication system. In the context of that toy model, I qualitatively develop the idea that different worlds may be more or less induction friendly (Section A.2). Instead of directly looking for the induction friendliness of the world, we will start by asking about the induction *unfriendliness* of a world, and doing that

requires quite a bit of setup. First, in Section A.3, I introduce toy frequencies that are like, but just not quite, the probabilities of individual (or combinations of) states in our toy worlds. The next, and crucial, step is to begin by considering the induction unfriendliness of a world in just the simple situation of focusing on just the regularities in the given system have the same antecedent kind. It is this situation that most closely parallels Shannon's theorem, and I re-motivate Shannon's three assumptions for the purpose of quantifying induction unfriendliness in Sections A.4, A.5, and A.6, respectively. In Section A.7, I derive the induction unfriendliness of a world when the given system is allowed to include regularities with differing antecedent kinds as being the conditional entropy of consequent states in the world on antecedent states. Finally, in Section A.8, I show how the induction friendliness of a world is approximately the MI between the antecedent and consequent states of the world.

As in Shannon, the success of the work in this section is not required for successfully arguing that induction friendliness, MI, and the BSA, should be brought together. The preceding sections should stand on their own. But I do think this exercise in reconstructing one of the original results of information theory is worthwhile. For one, it illustrates in significant detail *how* induction friendliness, MI, and the BSA, can be brought together. I also think it is emblematic of how the sorting out of the details of the best system competition should be done—namely, by carefully going through the needs and interests of scientific practice, and seeing what emerges from that.

A.1 THE TOY BSA

To begin, there are the local qualities, the basic kinds, that are distributed throughout the world. The toy-ness of our model appears most strongly in the assumption that the world is a discrete and finite string of the basic kinds. Any toy world may be described as a 'coin flip' world where there are only two kinds `0` and `1` (or `heads` and `tails` or `H` and `T` or whatever one prefers). Here is an example of a coin flip world:

$$w_{\mathrm{e}} = \texttt{10100101001010101111100101}$$

By default I assume that a toy world is a coin flip world. I will refer to any position in the world where a kind obtains a "state", as in "the first two states of w_{e} are `1` and `0`, respectively". When we are interested in more than two kinds I will say that the world is what it is "as expressed in L", where L consists in a set of kinds K and a function $T(-)$ from binary sequences to K. So, for an example language L_{e}, if

$K_e = \{\texttt{a}, \texttt{b}, \texttt{c}, \texttt{d}\}$, and T_e maps `10` to `a`, `01` to `b`, `00` to `c`, and `11` to `d`, we may say of our example world from above that

$$w_e =_{L_e} \texttt{aabbcaaaddabb}.$$

That is, the world w_e as expressed in L_e—or, with kinds K_e according to the translation T_e—is "`aabbcaaaddabb`".

To accompany our toy worlds is a toy conception of the BSA and induction. A system S, always to be paired with a language L, is characterized by a set of regularities of the form "If antecedent kind a obtains at any time t in the world, then the consequent kind c obtains at $t + 1$" for any $a, c \in K$. I will just write "$a \rightarrow c$" when there will be no loss of clarity. A best system competition for a world w scores pairs $S_i, L_i \in \mathcal{S} \times \mathcal{L}$ according to (usually) the balance of simplicity and strength S_i has with respect to w when both are expressed in L_i. The laws of world w are the regularities that appear in S_{best}, where $S_{\text{best}}, L_{\text{best}}$ is the highest scored of the competing system-language pairs.

Epistemologically, the talk of simplicity and strength is providing a particular, if rather vague, story about the inductive practices of scientists. To say that S, L is the (epistemologically) best system-language pair—whatever regularities appear in our best estimates of the laws and however we generalize from what limited data we have to the whole world—is to say that it is the product of our best inductive practices. Part of the aim of Chapter 5 was to step back from the standard picture of the best system being the simplest and strongest on balance. In the place of simplicity and strength, we would like to install induction friendliness, but that cannot be done without the arguments to come. For now, we will suppose for any S, L that it has been given to science from on high as the conclusion of some ideal inductive practice.

In the context of the BSA, we normally think of S, L as being better or worse for a particular world w. But for the moment we are being neutral on what makes S, L better or worse; all we know is that it is the best. It is helpful, then, to flip our usual thinking around to consider, for a given S, L, "How good is the world w?". Or, since S, L is the product of an ideal inductive practice: "How induction friendly is w?"

A.2 DEGREES OF INDUCTION (UN)FRIENDLINESS

We can start looking for an answer to the question of "How induction friendly is w?" by illustrating Hume's worry about the failure of constant conjunction in our toy worlds. The worry, at least on the surface,

is that we see constant conjunction until some time at which there is a failure of the pattern. Such a possible world would be like w_{h} below.[2]

$$w_{\mathrm{h}} = \overline{\mathtt{01}}|_t \mathtt{00}...$$

In w_{h} we see a repetition of `0` followed by `1` until the step from t to $t+1$ when the constant conjunction fails and a `0` is followed by another `0` instead of a `1`. Hume is not really concerned about what happens after that (hence the ellipsis)—what matters is just that the constant conjunction did fail, and nothing that happened prior to t assures us that we aren't in a world with such a failure.

We can motivate the possibility of there being more or less inductively friendly worlds by going beyond Hume and considering what the world looks like after that initial failure of the `01` pattern, as in the worlds described below.

$$\begin{aligned} w_{1\times\mathtt{00}} &= \overline{\mathtt{01}}|_t \mathtt{00}\overline{\mathtt{01}} \\ w_{2\times\mathtt{00}} &= \overline{\mathtt{01}}|_t \mathtt{000}\overline{\mathtt{01}} \\ &\vdots \\ w_{i\times\mathtt{00}} &= \overline{\mathtt{01}}|_t \overline{\mathtt{00}^{i}\mathtt{01}} \end{aligned}$$

Each of the $w_{i\times\mathtt{00}}$ worlds is consistent with w_{h}.[3] In $w_{1\times\mathtt{00}}$ the violation of the `01` pattern occurs only once. As failures of induction go, this presumably is not a bad one—yes there was a failure, but it is just the one and everywhere else our adoption of the $\mathtt{0} \to \mathtt{1}$ regularity will work so well. Similarly for $w_{2\times\mathtt{00}}$—clearly it is a `01` world with just a small wrinkle in the middle—but things are bit worse than they were in $w_{1\times\mathtt{00}}$ since the dominant regularity isn't quite as dominant. The inductive inference that we presumably would make from observing the world prior to t to thinking that `0` is always followed by `1` seems less and less successful as the period in which that regularity fails to apply gets larger.

So it seems as though the world looks less induction friendly as i becomes larger. But things shouldn't be quite so simple. As i grows to make the `00` pattern a significant fraction of the world, we might stop thinking that we're looking at a `01` patterned world with some

2 The subscript of vertical bar names the following point in time. An ellipsis indicates an indefinite but finite arbitrary string. An over-line indicates indefinite but finite repetitions of the over-lined sequence, or, when followed by a superscript, that superscript's number of repetitions.

3 Note that w_{h} actually describes a set of possible worlds on account of the openness of the ellipsis and number of initial repeated `01` pairs, and included in that set are the $w_{i\times\mathtt{00}}$ worlds. That is what I mean here by "consistent".

failure in the middle, but rather a world with a 01 regime, followed by a 00 regime, and then another 01 regime, and perhaps there is some threshold where that makes for a more induction friendly world than the one in the series of worlds before it. Further still, as i becomes large relative to the total size of the world, what we will see is not a 01 world with some mess in the middle, or a 01 to 00 to 01 world, but rather a 00 world with some mess on the ends, and that may present another inflection point in the plot (against i) of the induction friendliness of the $w_{i\times 00}$ worlds.

We may also consider the problem of gruesome predicates, the titular "new riddle of induction", raised in Goodman (1954): "Grue" is a predicate that "applies to all things examined before t just in case they are green [in appearance] but to other things just in case they are blue" in appearance (Goodman 1954, p. 74). Similarly, "bleen" applies to a thing just in case it appears blue when examined prior to t and green after that. Notably, anything that we can say about green and blue can also be said in terms of grue and bleen, and vice versa. Prior to time t, all the emeralds we have examined have *appeared* green, but, prior to t, appearing green is indistinguishable from appearing grue. While we might like to generalize from our examinations of emeralds to the conclusion that all emeralds are green, the evidence is just as good in favor of the incompatible alternative conclusion that all emeralds are grue.

The obvious appeal of treating this problem in the context of the (toy) BSA is that either (1) t is at or after the end of our finite world, and so green and grue are effectively indistinguishable, or (2) we have access to the facts after t, and so can say with confidence whether emeralds are actually green or grue (depending on whether they appear green or blue after t, respectively). In the case of (1), generalizing to all emeralds are green is not incompatible with generalizing to all emeralds are grue—the two regularities will either both be true or both be false in every possible world that ends at or before t—and so the problem goes away. In the case of (2): If emeralds do, in fact, persist in appearing green, then the employer of grue (and bleen) will be confronted with a world in which there is an abrupt shift from emeralds appearing grue to emeralds appearing bleen. When described using the language of grue and bleen, such a world seems less friendly to induction, but when described using the language of green and blue it seems perfectly induction friendly (with respect to emeralds) since emeralds persist in their greenness. Similarly, if the world is such that emeralds *do* change appearance from green to blue, then the world will seem more induction friendly if we describe it in terms of grue and bleen as opposed to green and blue.

The toy worlds that I have described cannot quite capture what is happening in Goodman's grue emerald thought experiment because the time-stamped translations from green and blue to grue and bleen (or vice versa) cannot straightforwardly be done with the time invariant translation functions $T(-)$. We can, though, appreciate that there should be a difference in induction friendliness between a world in which the same predicate always applies (as when the emerald always appears green) and a world with an initial segment of one kind occurring followed by some other kind occurring in the rest of the world (as when we find the emerald switches appearance from green to blue). We can likewise appreciate that the difference in induction friendliness should be perfectly reversed if we were to redescribe each world in such a way as to make the first world switch the prevailing kind (as when the emerald switches from grue to bleen) and the second world have a single prevailing kind (as when the emerald is always grue).

The time-stamping involved in the translations from green and blue to grue and bleen is not necessary for the induction friendliness of a world to change under redescription. Consider the following worlds (with some spacing added for clarity):

$$\begin{aligned} w_{\mathrm{a}} &= \texttt{10 01 10 01 10 01 10 01 10 01} \\ w_{\mathrm{b}} &= \texttt{110 011 110 001 110 001 100 011 100 001} \end{aligned}$$

The world w_{a} is clearly quite regular, but not in a way that can be captured by our restricted conception of what regularities may appear in a system because we can use just those concerned with what single state follows another single state. To correct this issue we could expand the sort of regularities that we allow for to include what pairs of states follow other pairs, but that is unnecessary. All we need to do is redescribe the world. Using the same language L_{e} introduced earlier—according to which each each `10` goes to `a` and `01` goes to `b`—we get that

$$w_{\mathrm{a}} =_{L_{\mathrm{e}}} \texttt{a b a b a b a b a b}.$$

Now we can say that w_a, redescribed using L, is perfectly regular; `a` is always followed by `b`, and `b` is always followed by `a` (at least if we ignore the end of the world not having any consequent state).

In contrast to w_{a}, w_{b} does not seem particularly regular. But what if we care about only the end points of triplets of states? Let L'_{e} be like L_{e} except that whenever the T_{e} of L_{e} takes a pair of states to some element of K_{e}, T'_{e} takes two triplets to the same element of K_{e} where the first and last states of the triplet and pair are the same, and one triplet as has a center state of `0` and the other a center state of `1`. For example,

while T_e maps 10 to a, T'_e maps both 100 and 110 to a. Expressing w_b in L'_e gets us that

$$w_b =_{L'_e} \texttt{a b a b a b a b a b}$$

So w_b is regular, at least when described in the right sort of way. This should come as no surprise. How induction friendly a world is depends on the language we use to describe the world. We could try to privilege as the "true" induction friendliness of the world whatever value is yielded by the "true" language for expressing the world, but in our toy worlds there is no such true language.

A.3 TOY FREQUENCIES

Let us return to our first example world

$$w_e =_{L_e} \texttt{aabbcaaaddabb}.$$

There are, straightforwardly, some actual relative frequencies for this world. Namely, there are the frequencies with which kinds obtain in the world. In w_e, expressed according to L_e, those frequencies are

$$\begin{aligned} f_{w_e,L_e}(\texttt{a}) &= 6/13 \\ f_{w_e,L_e}(\texttt{b}) &= 4/13 \\ f_{w_e,L_e}(\texttt{c}) &= 1/13 \\ f_{w_e,L_e}(\texttt{d}) &= 2/13. \end{aligned}$$

For convenience, I will drop the subscripts on f when they are clear in context.

There are also frequencies associated with seeing pairs of kinds obtaining. Continuing with our example,

$$\begin{aligned} f(\texttt{aa}) &= 1/4 \\ f(\texttt{ab}) &= 1/6 \\ f(\texttt{ac}) &= 0 \\ f(\texttt{ad}) &= 1/12 \end{aligned}$$

and so on for all 16 possible pairs. These frequencies are interesting in our toy worlds since they measure the frequency with which a corresponding regularity is the correct one to apply. For example, $f(\texttt{ab}) = 1/6$ indicates that the regularity $\texttt{a} \to \texttt{b}$ correctly characterizes

a sixth of all transitions from one state to the next. With this point in mind, I will call these "significance" frequencies.

In addition to the significance frequencies, it will be helpful to consider the frequency with which a regularity will be deployed successfully by someone with knowledge of the antecedent state but not the consequent state. For some $a, c \in K$, the "success" frequency $f_{w,L}(c|a)$ is the frequency with which c is the second kind in all state pairs that begin with a. In our example

$$\begin{aligned} f(\mathsf{a}|\mathsf{a}) &= 1/2 \\ f(\mathsf{b}|\mathsf{a}) &= 1/3 \\ f(\mathsf{c}|\mathsf{a}) &= 0 \\ f(\mathsf{d}|\mathsf{a}) &= 1/6 \end{aligned}$$

and so on.

These frequencies behave almost like regular probabilities. The sum over all $k \in K$ of $f(k)$ will equal one, as a regular probability would. We also have it that

$$f(c|a) \approx \frac{f(ac)}{f(a)}$$

where the failure to achieve equality of the two sides is due to the first and last states in a world not having antecedent or consequent states (respectively). However, as worlds become large, the difference from equality goes to zero. This near equality is reminiscent of the standard definition of conditional probability—the conditional probability $p(x|y)$ is the joint probability $p(x, y)$ divided by the probability $p(y)$—suggesting that success frequencies are analogous to conditional probabilities and significance frequencies to joint probabilities. The biggest failure of this analogy will be due to the fact that $f(ac) \neq f(ca)$ but $p(x, y) = p(y, x)$, and the rest will be due to the end effects already mentioned that become smaller as worlds become larger.

A.4 CONTINUITY

Suppose we have observed that some kind a ($\in K$) obtains at t. We would like to know what happens at $t + 1$. If our given system S contains just the single regularity $a \rightarrow c$, then we at least know what to expect. It of course could fail to be the case that in this particular instance c is what obtains at $t + 1$; S may be the best system, but that doesn't guarantee that the regularity will hold in every situation in which it is applicable (where a regularity is "applicable" at a time t

just in case its antecedent kind obtains at t).[4] What is at issue is the success frequency associated with the regularity: If $f(c|a) = 1$, then every a is followed by a c and the regularity $a \rightarrow c$ holds everywhere it is applicable. If $f(c|a) < 1$, then the regularity fails with a frequency of $1 - f(c|a)$ among the points in time when it is applicable. In general (for all $a, c \in K$), the use (when it is applicable) of a regularity $a \rightarrow c$ will be successful with frequency $f(c|a)$—this is precisely the conditional-probability-like "success" frequency introduced above.

Situated in this world where our best inductive practices recommend the adoption of the $a \rightarrow c$ regularity, what would we like to be true of $f(c|a)$? Clearly we would do best if $f(c|a) = 1$, as then every time we see an a we would correctly predict the appearance of a c. As the value of $f(c|a)$ falls away from 1, its successful use becomes less frequent, and, presumably, the induction friendliness of a world where $a \rightarrow c$ is truly the single best regularity to adopt goes down proportionally.

Of course, $a \rightarrow c$ need not be the only regularity that has been recommended by our best inductive practices. Continuing to restrict ourselves to the case where we have observed a, the set of applicable regularities (adopted on the recommendation of our best inductive practices or not) is comprised of the regularities $a \rightarrow c_i$ for all $c_i \in K$. Compare two worlds using the same language. If $f(c_i|a)$—which, since a is fixed for the time being, I will write as $f_a(c_i)$ for convenience and clarity—is lower in the first world than the second, then necessarily the success frequency of at least one of the other applicable regularities will be lower in the second world than in the first. In other words, if $f_{a,w,L}(c_i) < f_{a,w',L}(c_i)$ there exists a $c_j \in K$ such that $j \neq i$ and $f_{a,w,L}(c_j) > f_{a,w',L}(c_j)$). It is thus not immediately obvious how strongly, when considering the full set of applicable regularities, deviation from a success frequency of 1 impacts the induction friendliness of a world, since being further from 1 is less friendly, but ensures that there is at least one other regularity that is more successful.

Let H be the function that is meant to quantify the induction unfriendliness of a world with respect to the best system whose regularities all feature a as their antecedent kind. It seems clear, at the very least, that H is a function of the $f_a(c)$ for all $c \in K$. I will assume something slightly stronger (and make all the dependencies explicit in subscripts):

4 This is in conflict with the idea that laws should be universal. We can assume that the non-universal nature of these regularities is just part of the toy-ness of the toy model. It might also be that laws are not—and should not be expected to be—universal. This is the position of Braddon-Mitchell (2001) and Schrenk (2008, 2014), and, if correct, the non-universality of of these regularities is not a problem at all. In Chapter 3, I discuss these violations of universality and refer to them as "simple system exceptions".

CONTINUITY. $H_{a,w,L}$ is a continuous function of just the frequencies $f_{a,w,L}(c)$ for all $c \in K$.

CONTINUITY is stronger than what is immediately apparent in light of the preceding discussion on account of insisting that (1) H is continuous, and that (2) H is a function of *just* the success frequencies.

Defending (2) is a matter or reemphasizing the simplifying assumptions surrounding our toy model and what we are trying to quantify. To say that H is a function of *just* the success frequencies is a bit disingenuous since the success frequencies themselves depend on the world and choice of language. The only information about the world that is not contained in the success frequencies has to do with the ordering and number of states (and even then there is some information since success frequencies are concerned with pairs). We should not worry about H not exploiting that information because such large scale ordering information is not generally available to someone trying to exploit regularities arrived at by induction.

In defense of (1), the continuity of H: Suppose we have two worlds w and w'. Without loss of generality, suppose further that we have left them as being described in the basic coin-flip form—this allows us the convenient fact that in each world $f_a(\mathtt{1}) = 1 - f_a(\mathtt{0})$, so there is no question when there is a difference in the success frequencies between the two worlds about how the difference is distributed across the relevant regularities. Let $f_{a,w}(\mathtt{1}) = f_{a,w'}(\mathtt{1}) + \Delta$. We should take H to be continuous if, for every possible value of $f_{a,w}(\mathtt{1})$, in the limit as Δ goes to 0, the difference between H_w and $H_{w'}$ also goes to zero.

Now consider an arbitrary value of $f_{w,a}(\mathtt{1})$. Say, $f_{a,w}(\mathtt{1}) = .75$. Since the relevant success frequency is close to, but not, 1, this world w isn't ideal, but it's also not terrible. Similarly, for some small positive value of Δ, the world w' won't be too bad. With respect to just the $a \to \mathtt{1}$ regularity, w' is definitely more induction unfriendly than w (since $f_{a,w}(\mathtt{1}) > f_{a,w'}(\mathtt{1})$)—note though that CONTINUITY makes no demands on the direction of the difference. There will be some mitigation of the difference between the induction unfriendliness of the two worlds when we consider all the applicable regularities because $f_{a,w}(\mathtt{0}) < f_{a,w'}(\mathtt{0})$, and so, with respect to just the $a \to \mathtt{0}$ regularity, w' is better than w. As $f_{a,w'}(\mathtt{1})$ gets closer to .75 (i.e., as Δ goes to zero), the induction unfriendliness of the two worlds should also converge as they become more and more alike. And so it should be in general for any value of $f_{a,w}(\mathtt{1})$.

There are two salient values of $f_{a,w}(\mathtt{1})$ where one might reasonably have intuitions against the continuity of H. At $f_{a,w}(\mathtt{1}) = .5$, any difference in the success frequencies would make w' better. At $f_{a,w}(\mathtt{1}) = 1$,

any difference in the success frequencies would make w' worse. We might take the salience of these points to indicate their possessing *very*—i.e., discontinuously—high and low values of H, respectively. This is question begging, but perhaps not more so than the above appeals to intuition in favor of the continuity of H. Here then is an indirect argument for continuity via an attack on discontinuity. Discontinuity has the prima facie disadvantage of leaving us with a measure H that is not mathematically well behaved. Its advantage is that, as long as we are using H to quantify a cardinal ranking of induction unfriendliness, the discontinuity might better reflect the ranking we are trying to quantify. But this is only an apparent advantage. If there is discontinuity to be found in relation to induction unfriendliness, it will be ambiguous whether it is discontinuity in H or discontinuity in our response to H. As a parallel example, think of a full belief threshold in our degrees of belief. Our degrees of belief can take values in the continuum from 0 to 1, but the move from partial or no belief to full belief, if thresholded, will be discontinuous (as in, once my degree of belief in P passes .95 I will say that I have a full belief in P). Things may be similar for induction unfriendliness. H itself can be continuous (with all the mathematical advantages that brings), and any apparent discontinuity may be pushed from H itself to how we respond to H.

A.5 MONOTONICITY

To motivate the second assumption that we will make about H, consider the following scenario: You can choose to be a scientist in either world w or w'. L is the language of the best system-language pair in w, and L' the best in w'. The antecedent kind a that is the concern of H appears in and is treated the same way by both L and L'. No kind in K or K' fails to obtain in the respective worlds, and $|K| < |K'|$. Lastly, the L and L' are such that

(1) for all $c_i, c_j \in K$, $f_{a,w,L}(c_i) = f_{a,w,L}(c_j)$, and

(2) for all $c'_i, c'_j \in K'$, $f_{a,w',L'}(c'_i) = f_{a,w',L'}(c'_j)$.

Stated less formally: In two worlds w and w' where you have just observed a, the consequent kinds are all equally likely, and there are more possible consequent kinds in w' (as described by L'), than w (as described by L). In which world would you rather be? We've assumed that, relative to a particular regularity $a \to c$, a world is more friendly, or less induction unfriendly, the closer $f_a(c)$ is to 1. When we move to considering all the applicable regularities the issue is confused because, as the success frequency of one regularity goes down, the suc-

cess frequency of at least one other must go up. In this case we know exactly how all the success frequencies compare. Since the $f_{a,w,L}(c)$ are all equal, each is equal to $1/|K|$. Similarly in w' we have it that, for all $c' \in K'$, $f_{a,w',L'}(c') = 1/|K'|$. Thus the success frequency of every applicable regularity in w is strictly closer to 1 than the success frequency of any applicable regularity in w'. So w is preferable to—that is, more induction friendly than—w'. When the success frequencies of the applicable regularities are all equal, a world is less preferable—that is, more induction unfriendly—when there are more applicable regularities to consider.

We make the above explicit in our second assumption about H:

MONOTONICITY. $H_{a,w,L}$ is a a monotonically increasing function of $|K|$ when, for all $c_i, c_j \in K$, $f_{a,w,L}(c_i) = f_{a,w,S,T}(c_j)$.

Of the assumptions that we have made (and will make) about H, MONOTONICITY is the only one that influences the direction of changes in H relative to changes in the relevant success frequencies. It is also worth noting that MONOTONICITY makes H a tracker of a kind of ontological simplicity[5] (or lack thereof) by counting a world and associated system-language pair as being less friendly to induction precisely when more kinds are being used.

A.6 DECOMPOSITION

Consider a particular relation between two languages for a world that I will call *decomposition*. L' is a decomposition of L in w iff

(D1) $K = K'$ except that there are $n \geq 2$ kinds $c_1, c_2, ..., c_n \in K$ that do not appear in K', and $n + 1$ kinds $c'_0, c'_1, c'_2, ..., c'_n \in K'$ that do not appear in K, and

(D2) T and T' are such that $w_L = w_{L'}$ except, for each of the c_i among the $c_1, c_2, ..., c_n \in K$, wherever c_i obtains in w_L, the ordered pair (c'_0, c'_i) obtains in $w_{L'}$.

In our running example with language L_e and world w_e: The language L'_e is a decomposition of L_e in w_e when $K_e = \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}\}$, $K'_e = \{\mathtt{c}, \mathtt{d}, \mathtt{x}, \mathtt{y}, \mathtt{z}\}$, and, with spacing for clarity,

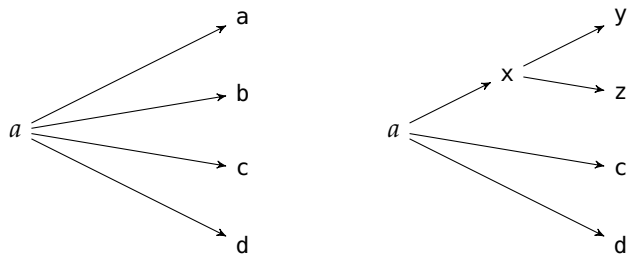
5 Schulte (2008) employs "ontological simplicity" in an analysis of the discovery of particle families. He characterizes ontological simplicity as encouraging the use of "as few ontological categories as possible"—for our purposes, this means having fewer kinds—and "as many particle families as possible that are disjoint, that is, categories whose boundaries do not overlap" (Schulte 2008, p. 307)

$w_e =_{L_e}$ a a b b c a a a d d a b b

$w_e =_{L'_e}$ xy xy xz xz c xy xy xy d d xy xz xz.

It will be convenient to be able to characterize a decomposition relation by saying what states decompose into what pairs of states. In our example we can characterize the decomposition relation by saying that "a decomposes to xy and b decomposes to xz" (in w, from L_e to L'_e).

Changing the language being used also changes how our given system S is expressed. If S, L contained the regularities $a \to$ a and $a \to$ b, S, L'_e should contain the regularities $a \to$ x, x $\to$ y, and x $\to$ z. This is illustrated in the picture below.



We can quickly say some things about the frequencies involving a, b, x, y, and z:

$$\begin{aligned} f_a(\mathsf{z}) &= f_a(\mathsf{a}) + f_a(\mathsf{b}) \\ f_a(\mathsf{a}) &= f_a(\mathsf{y}|\mathsf{x}) f_a(\mathsf{x}) \\ f_a(\mathsf{b}) &= f_a(\mathsf{z}|\mathsf{x}) f_a(\mathsf{x}). \end{aligned}$$

The success frequencies associated with the L_e kinds a and b are no different from chained success frequencies in L'_e of going from a to x and then from x to y or z. Effectively all that has happened is that we changed the names of the a and b kinds to xy and xz, respectively. It thus seems fair to assume that our move from L_e to L'_e makes no difference to the inductive unfriendliness of the world. In particular, the contribution to induction friendliness made by having the kinds a, b, and their associated regularities, in S, L_e should be the same as the contributions made by having x, y, z, and *their* associated regularities in S, L'_e. Similarly, the contributions made by the kinds c and d should be the same because they were unaffected by the decomposition.

The tricky part with this is that the measure of unfriendliness H that we have been building up is only defined for collections of regularities that have the same antecedent state, which is a rule we want to break

when considering the unfriendliness of S, L'. To get around this, we can decompose the regularities in S, L'_e: First we look at the H of the regularities taking us from a to x, c, and d. Then to that we add the H associated with the regularities taking us from x to y and z, but weighted by the frequency with which those regularities are applicable (conditional on a having preceded the x). This is our third assumption about H:

DECOMPOSITION. If L' is a decomposition of L in w where the kinds $k_1, ..., k_n \in K$ are decomposed into pairs of K' kinds $(k'_0, k'_1), ..., (k'_0, k'_n)$ and the kinds a and $k_{n+1}, ..., k_{n+m}$ appear in both K and K'. Then

$$H_{w,S,L}(w_{(c|a)}) = H_{w,S,L'}(w_{(c|a)}) + f_{a,w,L'}(k'_0) \times H_{w,S,L'}(w_{(c|a,k'_0)})$$

In the above, $w_{(c|a)}$ stands for the sequence of states in the world that are consequent to a and, similarly, $w_{(c|a,k'_0)}$ stands for the states consequent to the ordered pair of states a, k'_0.

Expanding on this to make the relevant (non-zero valued) frequencies explicit, we get

$$\begin{aligned} H_{w,S,L}(f_a(k_1), ..., f_a(k_{n+m})) = {} & H_{w,S,L'}(f_a(c'_0), f_a(k_1), ..., f_a(k_m)) \\ & + f_{a,w,L'}(k'_0) \times H_{a,w,L'}(f_a(k'_1|k'_0), ..., f_a(k'_n|k'_0)). \end{aligned}$$

And, spelled out in our running example, we get that

$$\begin{aligned} H_{w,S,L_e}(f_a(\mathsf{a}), ..., f_a(\mathsf{d})) = {} & H_{w,S,L'_e}(f_a(\mathsf{x}), f_a(\mathsf{c}), f_a(\mathsf{d})) \\ & + f_{a,w,L'_e}(\mathsf{x}) \times H_{w,S,L'_e}(f_a(\mathsf{y}|\mathsf{x}), f_a(\mathsf{z}|\mathsf{x})). \end{aligned}$$

A.7 INDUCTION UNFRIENDLINESS

Our three assumptions about H—CONTINUITY, MONOTONICITY, and DECOMPOSITION—are shown in Shannon (1948) to be satisfied uniquely by the function

$$H_{w,S,L}(w_{(c|a)}) = -k \sum_{c \in K} f_a(c) \log f_a(c)$$

where k is an arbitrary constant that determines the units of H (e.g. $k = 1/\log(2)$ gives units of bits). For a variety of reasons—most notably, its formal resemblance to the concept from statistical physics—H is sometimes called the *entropy* of $w_{(c|a)}$.

The entropy of $w_{(c|a)}$ is the amount of induction unfriendliness for a world associated with the regularities in the best system that have a as

their antecedent kind. In our running example, a could be any of a, b, c, or d. The best system S of world w expressed in L presumably contains regularities that feature all four kinds in the antecedent state position, and so we have all of $H(w_{(c|\mathtt{a})})$, $H(w_{(c|\mathtt{b})})$, $H(w_{(c|\mathtt{c})})$, and $H(w_{(c|\mathtt{d})})$, as relevant to the total unfriendliness of the world. Let each contribute in proportion to how frequently their respective regularities are applicable; that is, let the total unfriendliness of the world be the average over possible values of a of the unfriendliness measures $H(w_{(c|a)})$:

$$\begin{aligned} E_a[H(w_{(c|a)})] &= \sum_{a\in K} f(a)[-k\sum_{c\in K} f_a(c)\log f_a(c)] \\ &= -k\sum_{a\in K}\sum_{c\in K} f(a)f_a(c)\log f_a(c). \end{aligned}$$

Noting that $f_a(c) = f(c|a)$, and that $f(ac) = f(c|a)f(a)$, we can rewrite the above as what is known as the *conditional entropy* of consequent states given antecedent states

$$H_{w,S,L}(w_{(c)}|w_{(a)}) = -k\sum_{a\in K}\sum_{c\in K} f(ac)\log f(c|a).$$

If one is inclined to adopt some of the terminology of information theory, then this result has a very sensible reading: The induction unfriendliness of a world is equivalent to our uncertainty about the immediate future when we know the immediate past. If you are not so inclined, then the preceding argument can stand on its own, and perhaps offer a more palatable alternative reading for some other uses of information theoretic language.

A.8 INDUCTION FRIENDLINESS

Let us pause to take quick stock of what has been done so far. We are assured that S, L is the best system-language pair of world w, and have set out to answer the question "How induction friendly is w?". We have gotten as far as saying that the induction *un*friendliness of w is $H(w_{(c)}|w_{(a)})$, the conditional entropy of w's consequent states given the antecedent states.

One way of thinking about our measure of induction unfriendliness is that it tells us how bad things *still* are even after we are given the best system and the current (antecedent) state of the world. This suggests that there is a sort of induction unfriendliness to the world before we are given the best system. I say "sort of" here because there is no induction involved at that point—without the best system in hand, there are no generalizations around that have been made based on

past experience—just raw guessing about what kind is about to obtain. Conveniently, we already know how to measure this. Recall that the induction unfriendliness of the world for a particular given antecedent state a is

$$H_{w,S,L}(w_{(c|a)}) = -k \sum_{c \in K} f_a(c) \log f_a(c).$$

When we were looking for the total induction unfriendliness of the world above, we were going to know the antecedent state, we just didn't know which one it would be and so we took the average for all $a \in K$. In the current situation, anything we know about the antecedent state is irrelevant since we do not have a system on hand to exploit that knowledge. Thus the only applicable frequencies are the raw unconditional frequencies of the consequent states, and we get an "initial unfriendliness" of a world w and language L as the entropy simply of the consequent states:

$$H_{w,S,L}(w_{(c)}) = -k \sum_{c \in K} f(c) \log f(c).$$

Now what of induction friendliness? The system language pair S, L is assumed to be the best because it is the output of best inductive practices. We have worked out that $H(w_{(c)}|w_{(a)})$ is a measure of how unfriendly the world is *after* the best inductive practices recommend adopting S, L. And we have just determined that $H_{w,S,L}(w_{(c)})$ is a measure of how unfriendly the world is *before* the best inductive practices recommend adopting S, L. Let the induction friendliness of the world be how much less the world is unfriendly after implementing the best inductive practices. That is, let the induction friendliness for a world w given the best system-language pair S, L, be

$$I_{w,S,L}(w_{(a)}, w_{(c)}) = H_{w,S,L}(w_{(c)}) - H(w_{(c)}|w_{(a)}).$$

This is *almost* the MI between the antecedent and consequent states of w because, while MI is symmetric—i.e. it would be that $I(w_{(a)}, w_{(c)}) = I(w_{(c)}, w_{(a)})$—this will not generally be true for us because of the asymmetry of $f(ac)$. However, we can see that it is formally similar by ex-

ploiting the fact that $f(ac) = f(c|a)f(a)$ and $f(c) = \sum_a f(c|a)f(a)$ are approximately true (with only edge effects violating equality):

$$\begin{aligned} I_{w,S,L}(w_{(a)}, w_{(c)}) &= H_{w,S,L}(w_{(c)}) - H(w_{(c)}|w_{(a)}) \\ &= k\sum_{c\in K}\sum_{a\in K} f(ac)\log f(c|a) - k\sum_{c\in K} f(c)\log f(c) \\ &= k\sum_{c\in K}\sum_{a\in K} f(ac)\log f(c|a) \\ &\quad -k\sum_{c\in K}\sum_{a\in K} f(c|a)f(a)\log f(c) \\ &= k\sum_{c\in K}\sum_{a\in K} f(ac)[\log f(c|a) - \log f(c)] \\ &= k\sum_{c\in K}\sum_{a\in K} f(ac)\log \frac{f(ac)}{f(a)f(c)} \end{aligned}$$

while the MI of two random variables X and Y is

$$I(X;Y) = k\sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}.$$

Why should we think that there is more than mere formal similarity between these two measures? The kinds of systems that we allow for has an effect on what our measure of induction friendliness will look like. If we restrict ourselves to asymmetric regularities in our systems—as we have—then we should expect an asymmetric measure of induction friendliness. If, for example, we allow for symmetric systems that can treat $w_{(a)}$ and $w_{(c)}$ as random variables running alongside each other, and not just as slightly offset fragments of the same sequence, then it makes perfect sense to equate $f(ac)$ and $p(a,c)$ since there will no longer be a privileged ordering between the states of $w_{(a)}$ and $w_{(c)}$. In general, something like the following should be true:

INDUCTION FRIENDLINESS. The induction friendliness of a world w and given system-language pair S, L is the mutual information between w's parts, where what "w's parts" are is determined by S, L.